HMMA401 Statistique Computationnelle 2020/2021
Jean-Michel Marin

Vous devez envoyer **avant le jeudi 18 février 2021 à 17h00** un rapport écrit de 15 pages maximum (hors annexes) **au format pdf** à l'adresse suivante `jean-michel.marin@umontpellier.fr`.

Le rapport doit clairement indiquer quelle a été la contribution de chaque membre du projet. Le listing du code R utilisé doit figurer en annexe du rapport.

## Projet 4 Bayesian inference for a log-linear model

A standard approach to the analysis of associations (or dependencies) between *categorical* variables (that is, variables that take a finite number of values) is to use *log-linear models*. These models are special cases of generalized linear models connected to the Poisson distribution, and their name stems from the fact that they have traditionally been based on the logarithmic link function.
In such models, a sufficient statistic is the *contingency table*, which is a multiple-entry table made up of the cross-classified counts for the different categorical variables.
The **airquality** dataset was obtained from the New York State Department of Conservation (ozone data) and from the American National Weather Service (meteorological data).
This dataset involves two repeated measurements over 111 consecutive days, namely the mean ozone $u$ (in parts per billion) from 1 pm to 3 pm at Roosevelt Island, the maximum daily temperature $v$ (in degrees F) at La Guardia Airport, and, in addition, the month $w$ (coded from 5 for May to 9 for September). If we discretize the measurements $u$ and $v$ into dichotomous variables (using the empirical median as the cutting point), we obtain the following three-way contingency table of counts per combination of the three (discretized) factors:

```
           month  5  6  7  8  9
ozone    temp
[1,31]   [57,79] 17  4  2  5 18
         (79,97]  0  2  3  3  2
(31,168] [57,79]  6  1  0  3  1
         (79,97]  1  2 21 12  8
```

This contingency table thus has $5 \times 2 \times 2 = 20$ entries deduced from the number of categories of the three factors, among which some are zero because the corresponding combination of the three factors has not been observed in the study.

Each term in the table being an integer, it can then in principle be modeled as a Poisson variable. If we denote the counts by $\mathbf{y} = (y_1, \ldots, y_n)$, where $i = 1, \ldots, n$ is an arbitrary way of indexing the cells of the table, we can thus assume that $y_i \sim \mathcal{P}(\mu_i)$. Obviously, the likelihood

$$\ell(\boldsymbol{\mu}|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{y_i!} \mu_i^{y_i} \exp(-\mu_i)\,,$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, shows that the model is *saturated*, namely that no structure can be exhibited because there are as many parameters as there are observations. To exhibit any structure, we need to constrain the $\mu_i$'s and do so via a GLM whose covariate matrix $\mathbf{X}$ is directly derived from the contingency table itself.
When we express the mean parameters $\mu_i$ of a log-linear model as

$$\log(\mu_i) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\,,$$

the covariate vector $\mathbf{x}_i$ is indeed quite special in that it is made up *only* of indicators. The so-called *incidence matrix* $\mathbf{X}$ with rows equal to the $\mathbf{x}_i$'s is thus such that its elements are all zeros or ones. Given a contingency table, the choice of indicator variables to include in $\mathbf{x}_i$ can vary, depending on what is deemed (or found) to be an important relation between some categorical variables. For instance, suppose that there are three categorical variables, $u$, $v$, and $w$ as in airquality, and that $u$ takes $I$ values, $v$ takes $J$ values, and $w$ takes $K$ values. If we only include the indicators for the values of the three categorical variables in $X$, we have

$$\log(\mu_\tau) = \sum_{b=1}^{I} \beta_b^u \mathbb{I}_b(u_\tau) + \sum_{b=1}^{J} \beta_b^v \mathbb{I}_b(v_\tau) + \sum_{b=1}^{K} \beta_b^w \mathbb{I}_b(w_\tau)\,;$$

that is, ($1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$),

$$\log(\mu_{l(i,j,k)}) = \beta_i^u + \beta_j^v + \beta_k^w$$

($1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$), where $l(i,j,k)$ corresponds to the index of the $(i,j,k)$ entry in the table, namely the case when $u = i$, $v = j$, and $w = k$. Similarly, the saturated log-linear model corresponds to the use of one indicator per entry of the table; that is,

$$\log(\mu_{l(i,j,k)}) = \beta_{ijk}^{uvw}$$

($1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$).
For comparison reasons that will very soon be apparent, and by analogy with analysis of variance (ANOVA) conventions, we can also over-parameterize this representation as

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw}\,, \tag{1}$$

where $\lambda$ appears as the overall or reference average effect, $\lambda_i^u$ appears as the marginal discrepancy (against the reference effect $\lambda$) when $u = i$, $\lambda_{ij}^{uv}$ as the interaction discrepancy (against the added effects $\lambda + \lambda_i^u + \lambda_j^v$) when $(u,v) = (i,j)$, etc.

Using the representation (1) is quite convenient because it allows a straightforward parameterization of the nonsaturated models, which then appear as submodels of (1) where some groups of parameters are null. For example,

(i) if both categorical variables $v$ and $w$ are irrelevant, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u\,;$$

(ii) if all three categorical variables are mutually independent, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w\,;$$

(iii) if $u$ and $v$ are associated but are both independent of $w$, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv}\,;$$

(iv) if $u$ and $v$ are conditionally independent given $w$, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ik}^{uw} + \lambda_{jk}^{vw}\,; \quad \text{and}$$

(v) if there is no three-factor interaction, then

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw}\,,$$

which appears as the most complete submodel (or as the global model if the saturated model is not considered at all).

As stressed above, the representation (1) is not identifiable. Although the following is not strictly necessary from a Bayesian point of view (since the Bayesian approach can handle nonidentifiable settings and still estimate properly identifiable quantities), it is customary to impose identifiability constraints on the parameters as in the ANOVA model. A common convention is to set to zero the parameters corresponding to the first category of each variable, which is equivalent to removing the indicator (or *dummy variable*) of the first category for each variable (or group of variables). For instance, for a $2 \times 2$ contingency table with two variables $u$ and $v$, both having two categories, say 1 and 2, the constraint could be

$$\lambda_1^u = \lambda_1^v = \lambda_{11}^{uv} = \lambda_{12}^{uv} = \lambda_{21}^{uv} = 0\,.$$

For notational convenience, we assume below that $\boldsymbol{\beta}$ is the vector of the parameters once the identifiability constraint has been applied and that $\mathbf{X}$ is the indicator matrix with the corresponding columns removed.
The choice of a prior distribution for log-linear models is open to debate, a good choice is the following prior distribution

$$\boldsymbol{\beta}|\mathbf{X} \sim \mathcal{N}\left(0_p, n(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right)$$

where $\mathbf{X}$ is an $n \times p$ matrix.

**1** Give the density of the posterior distribution up to its normalizing contants. This is a non-standard density. This is our target density.

**2** We consider the most general nonsaturated model, case (v). How many parameters does this model contain? Create the matrix $\mathbf{X}$ and the associated vector $y$ with the following correspondances between the columns of $\mathbf{X}$ and the levels of the factors.

```
X1: intercept
X2: ozo2
X3: temp2
X4: mon2
X5: mon3
X6: mon4
X7: mon5
X8: ozo2:temp2
X9: ozo2:mon2
X10: ozo2:mon3
X11: ozo2:mon4
X12: ozo2:mon5
X13: temp2:mon2
X14: temp2:mon3
X15: temp2:mon4
X16: temp2:mon5
```

**3** Use an Hastings-Metropolis algorithm based on a multivariate Gaussian random walk to get some realizations coming approximately from the target posterior distribution.

As starting value, one can use the maximum likelihood estimation $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and as scale matrix the estimated covariance matrix $\hat{\Sigma}$ of the maximum likelihood estimator.

```
mod <- summary(glm(y~-1+X,family=poisson()))
```

provides $\hat{\boldsymbol{\beta}}$ in mod$coeff[,1] and $\hat{\Sigma}$ in mod$cov.unscaled.

**4** Compare the maximum likelihood and the bayesian estimates of $\boldsymbol{\beta}$.